Data first: turning the digital library "inside out"

Stephen Davison
sdavison@caltech.edu

Robert Doiel
rsdoiel@caltech.edu

Outside-in and Inside-out

Lorcan Dempsey's weblog, January 11, 2010
http://orweblog.oclc.org/Outside-in-and-inside-out/

"Think ... of a distinction between outside-in resources, where the library is buying or licensing materials from external providers and making them accessible to a local audience (e.g. books and journals), and 'inside-out' resources which may be unique to an institution (e.g. digitized images, research materials) where the audience is both local and external. Thinking about an external non-institutional audience, and how to reach it, poses some new questions for the library."

Caltech's Inside-out Library

- Institutional repository
  - **EPrints**
- ETDs
  - **EPrints**
- Research data repository
  - **Invenio**

Caltech's Inside-out Library (continued)

- Digital collections
  - **Islandora**
  - all formats, including born-digital data

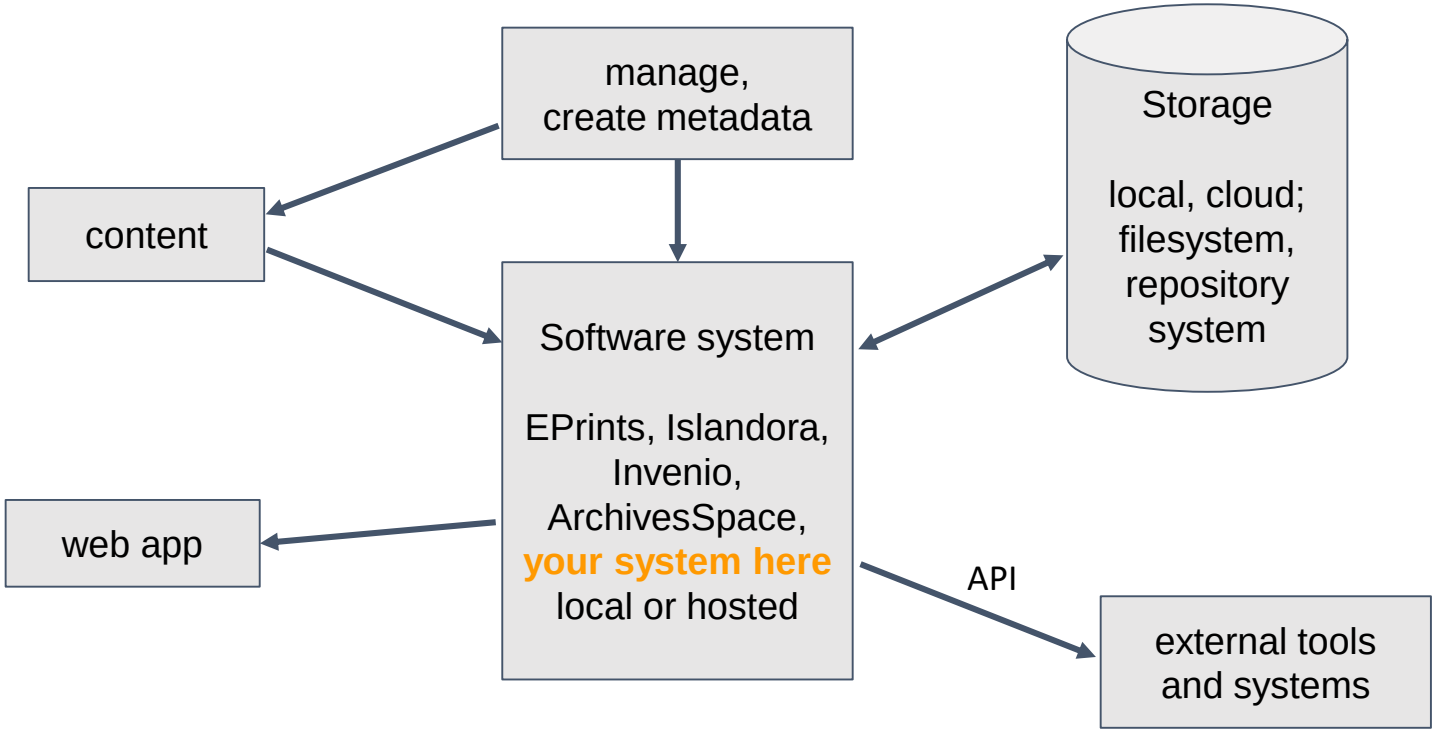
- Archives management
  - **ArchivesSpace**

Infrastructure issues

- Migrating to new software versions
- Development of similar capabilities across different systems
- Developing services against a variety of APIs
- Aggregation challenges

Caltech Directory

Caltech Library


manage,
create metadata
content
Software system
EPrints, Islandora,
Invenio,
ArchivesSpace,
your system here
local or hosted
Storage
local, cloud;
filesystem,
repository
system
web app
API
external tools
and systems

Objectives (things we need to do/support)

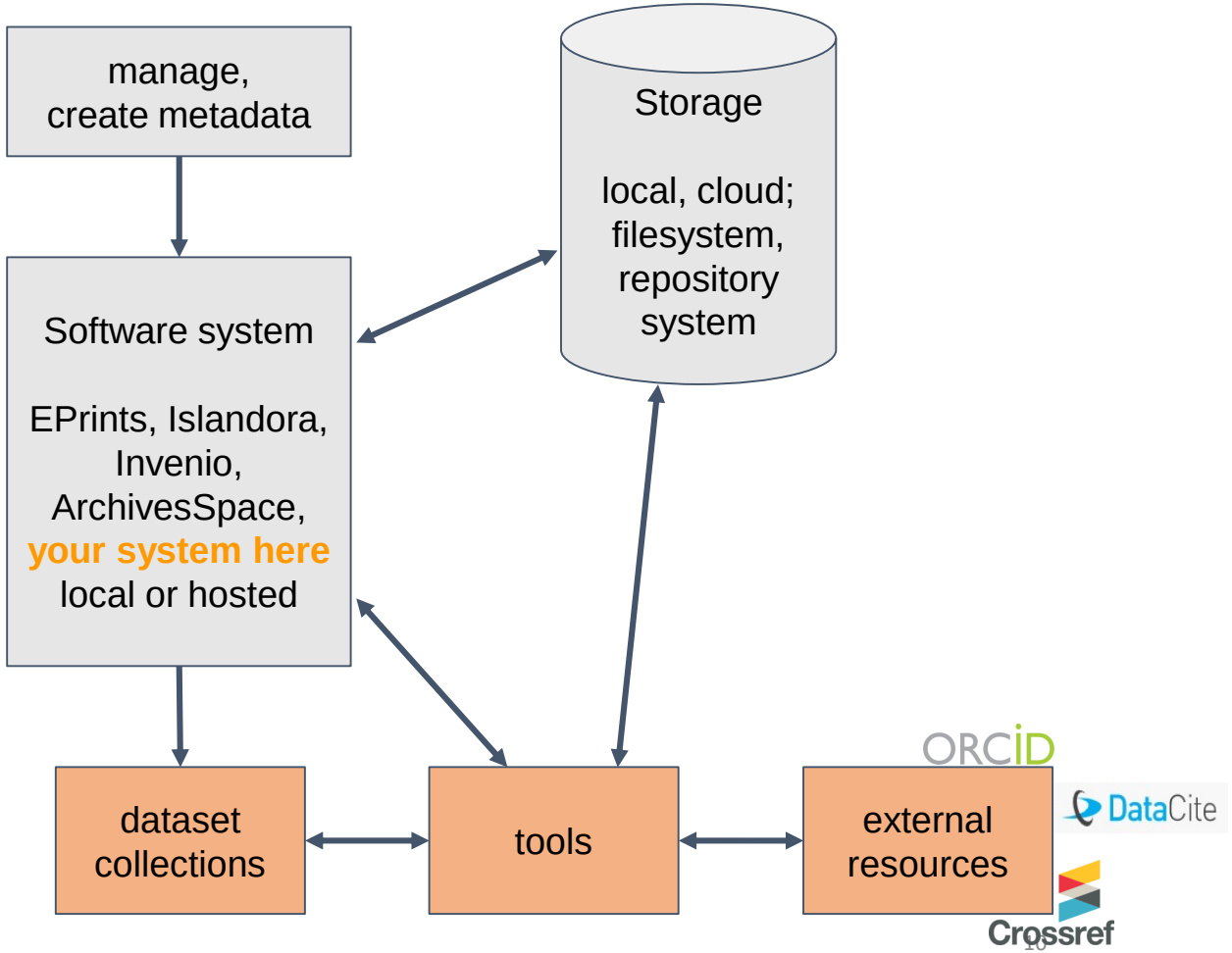
- Provide access to metadata in a variety of ways (feeds)
- Embed metadata and content in diverse contexts (publication)
- Search, retrieve, display
- Enrich, recombine, analyze

Software strategies

- Leverage APIs (working at the edge)
- Working with copies (of metadata primarily)
- Continuous migration (e.g. daily download)
- Simplicity
- Generic tools

Composability and malleability

- Composable software
  - Easily modifiable
  - No arbitrary restrictions
- Malleable software
  - Modular components


manage,
create metadata
Storage
local, cloud;
filesystem,
repository
system
Software system
EPrints, Islandora,
Invenio,
ArchivesSpace,
your system here
local or hosted
dataset
collections
tools
external
resources
ORCID
DataCite
Crossref

Tools

dataset

- Tools for organizing JSON objects as collections
  - on disc or in the cloud

ames

- Automated Metadata Service
- Manages metadata from a variety of sources
  - Metadata harvesting
  - CodeMeta management and updating
  - Citation alerts
  - Metadata checks and updates
  - Caltech repository reports

Strategic advantages

- Less need to develop WITHIN software/repository silos
- Preservation (“lots of copies keeps stuff safe”)
- Ease of manipulation
  - Filtering (of records, of fields)
  - Repurposing
  - Combination, distribution

Welcome to Caltech Library's aggregated feeds

Overview

Caltech Library operates a number of repositories including repositories for *thesis and dissertations*, *journal articles and publications* and *scientific data, models and software*. This site seeks to bring that content together in a manner to encourage re-use in websites and projects.

Including feed content is a easy as using one of our *Widgets* and pasting the HTML/JavaScript into your web page.

Aggregations

Content is aggregated around

- *recent* - Recent additions to CaltechAUTHORS and CaltechDATA
- *groups* - A curated list of organizations associated with articles, publications and data records in either *CaltechAUTHORS* or *CaltechDATA* repositories.
- *people* - A curated list of people of Caltech people with records associated in one of Caltech Library's repositories

Inside "groups" and "people" you'll find an alphabetical listing of links to individual groups and people. Clicking through will bring you to that group or person's profile page showing links to that various individuals feeds like articles, publications, data, models and software. You'll also find a link to "recent". Recent is like the groups, people and person feeds but with listings capped at the 25 most recent items. All feeds are sorted in descending publication date order.

Formats

Each feed is available in the following formats - HTML, HTML include, Markdown, BibTeX, JSON and RSS. Each format is indicated by it's file extension.

Example: Identity

- IDs and roles vary from system to system
- Roles and affiliations change over time
- Systems may or may not have “authority control”
- Options: normalize data, or use crosswalks

feeds.library.caltech.edu/people/Mead-C-A/

Mead, Carver

Links and identifiers

- Caltech Archives *profile*
- LCNAF *n79003041*
- ISNI *0000 0001 1756 9138*
- SNAC *w6c83dbk*
- Wikidata *Q62910*

Title

Gordon and Betty Moore Professor of Engineering and Applied Science, Emeritus

Division

Engineering and Applied Science Division

Biography

B.S., Caltech, 1956; M.S., 1957; Ph.D., 1960; D.Sc.h.c., University of Lund (Sweden); D.h.c., University of Southern California. Instructor in Electrical Engineering, Caltech, 1958-59; Assistant Professor, 1959-62; Associate Professor, 1962-67; Professor, 1967-77; Professor of Computer Science and Electrical Engineering, 1977-80; Moore Professor of Computer Science, 1980-92; Moore Professor of Engineering and Applied Science, 1992-99; Moore Professor Emeritus, 1999-.

(also available *recent 25* feeds)

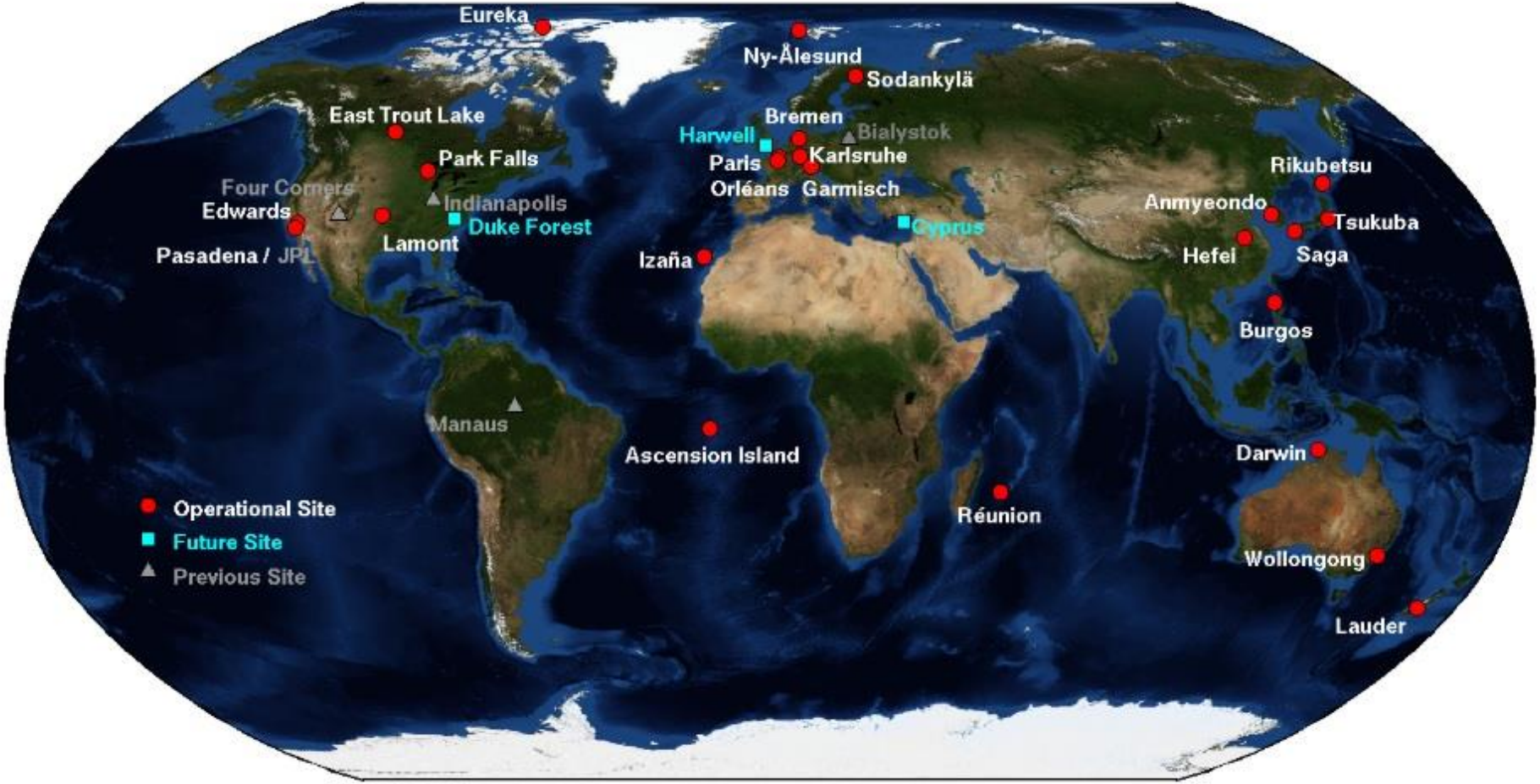
CaltechTHESIS

Mead, Carver (1960) **Transistor switching analysis** (Dissertation (Ph.D)), California Institute of Technology *https://resolver.caltech.edu/CaltechETD:etd-05182006-084112*

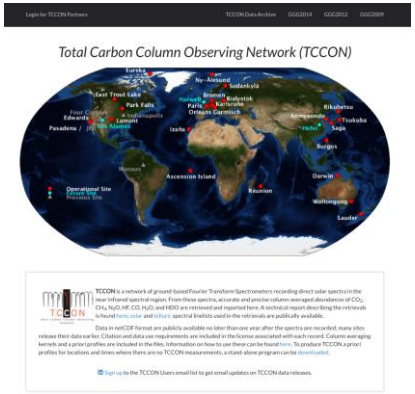
CaltechAUTHORS

- Combined (246) *HTML*, *HTML Include*, *Markdown*, *BibTeX*, *JSON*, *RSS*
- Article(s) (142) *HTML*, *HTML Include*, *Markdown*, *BibTeX*, *JSON*, *RSS*
- Book(s) (3) *HTML*, *HTML Include*, *Markdown*, *BibTeX*, *JSON*, *RSS*
- Book Section(s) (83) *HTML*, *HTML Include*, *Markdown*, *BibTeX*, *JSON*, *RSS*
- Conference Item(s) (1) *HTML*, *HTML Include*, *Markdown*, *BibTeX*, *JSON*, *RSS*
- Monograph(s) (17) *HTML*, *HTML Include*, *Markdown*, *BibTeX*, *JSON*, *RSS*

Total Carbon Column Observing Network (TCCON)



Departmental Server at Caltech



Data Files

Metadata

California Institute of Technology
Research Data Repository

Update metadata
Dates
Version
Information
Usage Information
Update files
Netcdf data

caltechdata_api

Monthly automatic updates, new data version or new site

DataCite
FIND, ACCESS, AND REUSE DATA

https://github.com/caltechlibrary/caltechdata_api
https://github.com/inveniosoftware/datacite

Leveraging DataCite to publish video

We've developed a JavaScript-based solution to embed video content from CaltechDATA into any web site. You only need to know the DOI of the CaltechDATA record. Paste the following code into your website:

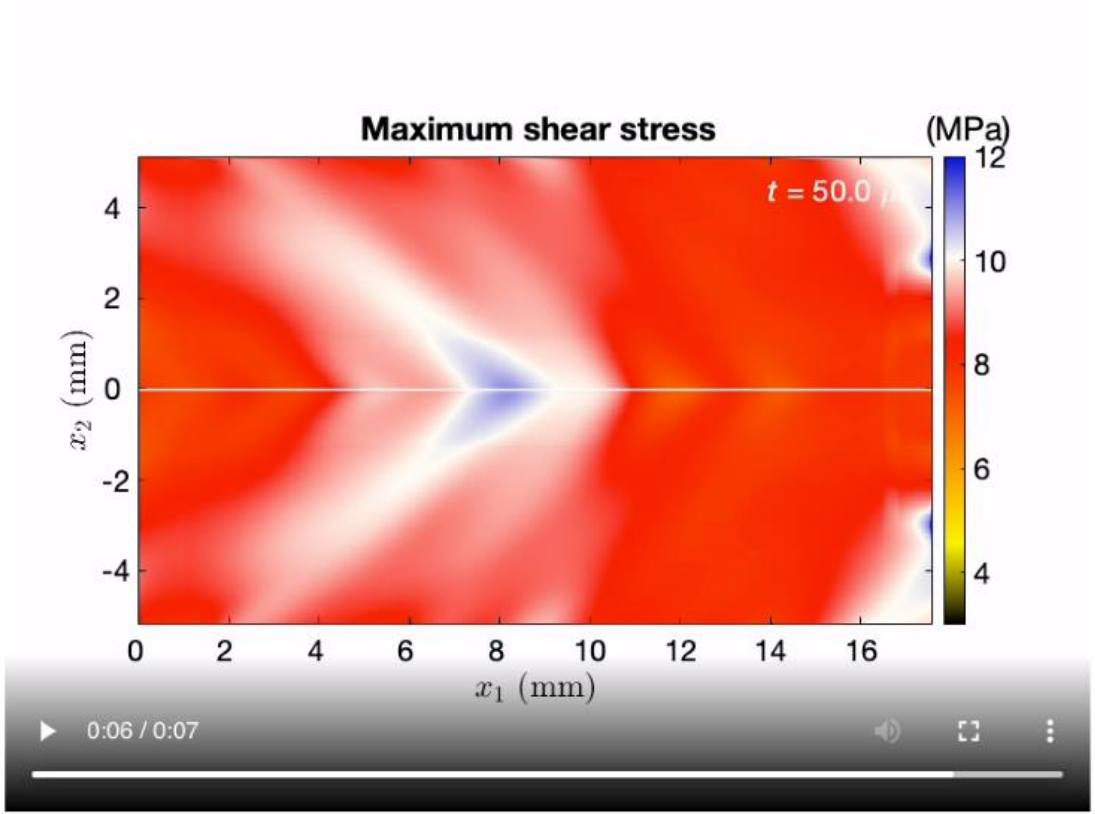
```
<div id="videodiv1"></div>
    <script src="https://feeds.library.caltech.edu/scripts/CL.js"></script>
    <script>
let div = document.getElementById("videodiv1"),
doi = '10.22002/D1.1278',
item_no = 0;
CL.doi_video_player(div,doi,item_no);
    </script>
```

Our CL.js function is general and will work with any DOI where the content provider has provided media information to DataCite. The underlying viewer is video.js, an open source javascript viewer.

Cell Atlas

- Proposal for a new publication
- Content to be hosted in CaltechDATA
- Publication via R bookdown package
  - Simple, flexible
  - Can include interactive apps
- Example: Embedded content (e.g. video) from CaltechDATA repository

skin"). This means that their cytoplasm is enclosed by a single membrane. At the resolution of this image, the membrane looks like a single dark line, but remember that it's really a bilayer, as you'll see in some later images.



Takeaways

- Complexity should reside in the data, not the system
- Value lies in the data, not the software
- Build at the edges
- Build tools, not systems